



TITLE:

R.A. Fisherにおける非正則推定論 (統計的漸近理論とその応用)

AUTHOR(S):

竹内, 啓

CITATION:

竹内, 啓. R.A. Fisherにおける非正則推定論(統計的漸近理論とその応用).
数理解析研究所講究録 1983, 507: 60-74

ISSUE DATE:

1983-12

URL:

<http://hdl.handle.net/2433/103759>

RIGHT:

R. A. Fisher における非正則推定論

竹内 啓

(Kei Takeuchi)

R. A. Fisher が統計的推定論に関連して、十分性、一
 致性、漸近有効性等の概念を導入して、正則な場合における
 最尤推定量の漸近有効性を主張したことはよく知られている。
 これは数学的には厳密な証明ではなかったとしても、その後
 多くの人々によって数学的に十分厳密な結果が確立されるに
 至る過程の出発点になった。また彼は推定量を用いることに
 よる情報量損失の概念を用いて、これまた正則な場合には、
 一次の漸近有効性を持つ推定量の中で、最尤推定量が漸近的
 情報量損失を最小にすること、という意味で2次の漸近有
 効性を持つことを示した。この点には後に C. R. Rao によっ
 て改め取上げられ、そして最近になって Pfanzagl,
 Ghosh, Takeuchi-Akashina による推定量の高次漸
 近有効性の理論として、完成された。このように Fisher は
 特に比較的初期の2論文:

Mathematical Foundations of Theoretical Statistics

1922 年, および

Theory of Statistical Estimation 1925 年

において、50年以上も後に完成されるまで、多くの理論的

成果を予見しているか、その中で破けた正則条件がたつ立
たない場合に何が起るであろうかについても、多くの結果
を予想を与えている。ここではそれを取り上げて検討するこ
とにしよう。

最初に推定理論における正則条件とは何を意味するかにつ
いてのべておこう。 $X_1, X_2, \dots, X_n, \dots$ を観測値の系列とす
るとき、それは通常次のことを意味する。

- 1) $X_1, X_2, \dots, X_n, \dots$ は互いに独立に母数 $\theta (\in \mathcal{H} \subset \mathbb{R}^p)$
をふくむ分布に従う。
- 2) 分布は密度関数 $f(x, \theta)$ を持つ。
- 3) 分布の台 support: $S = \{x \mid f(x, \theta) > 0\}$ は θ に無関
係に定めらる。
- 4) $f(x, \theta)$ は θ に関して連続微分可能。
- 5) Fisher 情報量行列

$$I(\theta) = E_{\theta} \left\{ \frac{\partial}{\partial \theta} \log f(X, \theta) \frac{\partial}{\partial \theta} \log f(X, \theta) \right\}$$

が有限。

このとき $\hat{\theta}_n$ を (X_1, \dots, X_n) にもとづく最尤推定量とす
ると、 $\sqrt{n}(\hat{\theta}_n - \theta)$ は n が大きくなるとき、漸近的に正規分布
 $N(0, I^{-1})$ に従い、かつそれが一次の漸近有効性 (定義はこ

こぞのべない)を持つことが知られている。そして更に $f(x, \theta)$ の θ に関する高次の連続微分可能性と $\partial^j \log f(x, \theta) / \partial \theta^j$ のモーメントの存在を仮定すると、適当に偏りを補正した最尤推定量が、2次ないし3次の漸近有効性を持つことが証明される (Akahira-Takenuchi [1] 参照)。

そこでこれらの条件のうちの一つ、或いは2つ以上がたり立たないとき、推定量の漸近的性質について、どういふことが起こり得るかをしるべることが、単に病理的のケースに対する意味というよりも、これらの正則条件の意義を明確にするために重要になる。

Fisher は上記の1922年の長大な論文の第8節で X_1, \dots, X_n が次のようなガンマ分布に従う場合を取り上げている。

$$f(x, \mu, \nu, \alpha) = \frac{1}{\alpha \Gamma(\nu)} \left(\frac{x-\mu}{\alpha} \right)^{\nu-1} e^{-(x-\mu)/\alpha} : x > \mu$$

Fisher はここで、母数の推定法として、モーメント法と最尤法の比較を行っている。

まず α, ν がともに既知の場合、モーメント法による μ の推定量は

$$\hat{\mu} = \bar{X} - \alpha \nu$$

であり、その分散は $\alpha^2 \nu / n$ となる。

他方 $L = \sum \log f(X_i)$ として

$$\frac{\partial^2 L}{\partial \mu^2} = -(\nu-1) \sum \frac{1}{(X_i - \mu)^2}$$

の期待値を計算する、 $\nu > 2$ とする。

$$nI(\mu) = -E\left(\frac{\partial^2 L}{\partial \mu^2}\right) = \frac{n}{\alpha^2(\nu-2)}$$

となるから、最尤推定量の漸近分散は $\alpha^2(\nu-2)/n$ となり、
従ってモーメント法の漸近効率 $(\nu-2)/\nu$ となる。

しかしこの議論にはやや問題がある。というのはここでは
上記の条件 3) が破られてゐるから、最尤推定量の漸近分布
に関する議論をそのまま適用することはできないのである。

しかしこの場合結論自体は正しいので、このことをもう少し
くわしく考察しよう。まず

$$\frac{\partial L}{\partial \mu} = -(\nu-1) \sum \frac{1}{(X_i - \mu)} + \frac{n}{\alpha}$$

であつて、 $\mu \rightarrow \min X_i$ のとき、 $\partial L / \partial \mu \rightarrow -\infty$ 、 $\mu \rightarrow -\infty$

のとき $\partial L / \partial \mu \rightarrow n/\alpha$ 、また上記の式から $\nu > 1$ ならば

$\partial^2 L / \partial \mu^2 < 0$ であるから、 $\partial L / \partial \mu = 0$ について唯一の解

$\mu^* < \min X_i$ を持ち、これが最尤推定量となる。そうし

て $\sqrt{n}(\mu^* - \mu)$ は漸近的に平均 0、分散 $1/I(\mu)$ の正規分布

に従うことは、通常の正則な場合と同様に、 $\partial L / \partial \mu$ をテラ
 う一層開くことにより証明できる。次にこれが一次の漸
 近有効性を持つことは、仮説 $\mu = \mu_0$ 。を対立仮説 $\mu = \mu_0$
 $+ t/\sqrt{n}$ に対して検定するとき、 $\hat{\mu}^*$ にもとづく検定の検出
 力が、漸近的に最強力検定の検出力に一致することを示せば
 よい。

$\nu \leq 2$ の場合については、Fisher は次のように言ってい
 る。「 $\nu = 2$ のときには、位置母数について標本平均の効率は
 0 になる。それは分布の端で密度関数が x 軸と正の角をたし
 ているからである。」と述べている。このことの意味づけにつ
 いては後に他の例題との関連で述べているが、彼は $\nu < 2$ の
 場合についても「標本平均の効率が 0 になる」とことは十分
 気づいていたと思われる。

実際 $\nu = 2$ の場合には最尤推定量 $\hat{\mu}_n^*$ と $\sqrt{n \log n} \times$
 $(\hat{\mu}_n - \mu)$ が漸近的に正規分布に従うことが示されるので、そ
 の漸近分散のオーダーは $(n \log n)^{-1}$ になり、 $\hat{\mu}$ の分散のオ
 ーダーより小さくなるのである。それ $\nu < 2$ の場合には
 $\hat{\mu} = \min X_i - c_n$ という形のものを作ると、その分散のオ
 ーダーは $n^{-2/\nu}$ となるので、 n^{-1} より小さくなるのである。
 更にここで言う一つ興味あることは、 $\nu = 2$ のときには最尤
 推定量は漸近有効性を持つことが知られるが (Woodroffe [4])

$\nu < 2$ の場合には、 $\hat{\mu}$ を持つ確率 1 という事である。従って $\nu < 2$ の場合について、Fisher がこれからのべきことを述べたわけ、その場合には最尤推定量の漸近的性質について明確なことがいえないということを感じたためであるかもしれない。

そこで $\nu < 2$ の場合についてもう少し考えてみよう。まず $\nu > 1$ ならば、上記の議論により、最尤推定量 $\hat{\mu}^*$ が、尤度方程式 $\partial L / \partial \mu = 0$ の一義的解として求まらぬことがわかる。そこで $\partial L / \partial \mu$ を展開すると

$$\frac{\partial L}{\partial \mu} \Big|_{\hat{\mu}^*} = 0 = \frac{\partial L}{\partial \mu} \Big|_{\mu} + \frac{\partial^2 L}{\partial \mu^2} \Big|_{\mu=\mu^0} (\hat{\mu}^* - \mu)$$

となる。ただしここで μ^0 は μ と $\hat{\mu}^*$ の間の値である。

この場合でも、 $\hat{\mu}^*$ が一貫性を持つことは明らかなので、 μ^0 を μ と置きかえて

$$\hat{\mu}^* - \mu \sim - \left(\frac{\partial L}{\partial \mu} \right) / \left(\frac{\partial^2 L}{\partial \mu^2} \right)$$

とすることが出来る。ここで $\nu < 2$ のときは

$$\frac{\partial L}{\partial \mu} = \sum \left(- \frac{\nu-1}{(x_i - \mu)} + \frac{1}{\alpha} \right)$$

の右項について、分散が無限大となるから、中心極限定理は $\partial L / \partial \mu$ について成立しない。そして $n^{-1/\nu} (\partial L / \partial \mu)$ の分布が

漸近的に指数 ν の stable law に収束することゝが証明される。
 また分母についても大数法則が成立せず、 $n^{-\frac{2}{\nu}}(\partial^2 L / \partial \mu^2)$ には
 ある確率分布を持つ。定数に確率収束しない。従って

$$n^{\frac{1}{\nu}}(\hat{\mu} - \mu) \sim \frac{n^{-\frac{1}{\nu}}(\partial L / \partial \mu)}{n^{-\frac{2}{\nu}}(\partial^2 L / \partial \mu^2)} \quad (*)$$

は、正規分布と異なる、非常に複雑な分布を持つ。

他方仮説 $\mu = \mu_0$ と対立仮説 $\mu = \mu_0 + t n^{-\frac{1}{\nu}}$ に対して検定
 するときの最強力検定の棄却域は $t > 0$ のとき、

$$\sum \log \frac{f(X_i, \mu_0 + t n^{-\frac{1}{\nu}})}{f(X_i, \mu_0)} \sim t n^{-\frac{1}{\nu}} \frac{\partial L}{\partial \mu} \Big|_{\mu_0} > c$$

で与えられる、これは最大推定量から得られる棄却域 $\hat{\mu} < \mu_0 + c n^{-\frac{1}{\nu}}$ とは、漸近的に一致しない。よって $(*)$ 式の右辺の
 分母が定数に確率収束しない、ことに注意するものであることを明
 しかねばならない。 $t < 0$ のときは最強力検定の棄却域は

$$\min X_i < \mu_0 + t n^{-\frac{1}{\nu}} \text{ および } \sum \log \frac{f(X_i, \mu_0 + t n^{-\frac{1}{\nu}})}{f(X_i, \mu_0)} > c$$

という 2 つの範囲から構成される、これも最大推定量から得ら
 れる棄却域には一致しない。

このまうな場合には、一様に (一次の) 有効性を持つ推定
 量は存在しない。

$\nu < 1$ のときは $\partial L / \partial \mu$ はつねに単調増加関数になるから、

最尤推定量は μ の可能な範囲の上限に一致し、 $\hat{\mu}^* = \min X_i$ となる。こうして $\nu = 1$ の場合には $\min X_i$ は十分統計量となるから、 $\hat{\mu}^*$ は(任意の次数の)漸近有効性を持つ。 $\nu < 1$ のときには $\hat{\mu}^*$ は漸近有効性を持たないがその収束のorderは $n^{-1/\nu}$ になるから、 $\hat{\mu}$ の漸近効率の0になることには変りはない。

次に ν が既知で、 μ, α がともに未知の場合には、モーメント法による推定量は、2次の標本モーメントを M_2 と表すと、 $M_2 = \hat{\alpha}^2 \nu$ となるから、 $\hat{\alpha} = (M_2/\nu)^{1/2}$ となり、その漸近分散は

$$(\nu+3)\alpha^2/2\nu n$$

となる。他方情報量行列は

$$n \begin{pmatrix} 1/\alpha^2(\nu-2) & 1/\alpha^2 \\ 1/\alpha^2 & \nu/\alpha^2 \end{pmatrix}$$

となるから、 α の最尤推定量の漸近分散は $\alpha^2/2n$ となり、従ってモーメント法による推定量の漸近効率は $\nu/(\nu+3)$ となる。この場合も、 $\nu > 2$ ならば、正則条件3)が満たされているにもかかわらず、このように結論が正しいことは証明できる。

ところでFisherは $\nu \leq 2$ の場合でも α の最尤推定量の漸

近分散が $\alpha^2/2n$ で与えられるようにのべているが、それは誤りである。 ν, μ がともに既知のとき、 α の最小分散不偏推定量は $\hat{\alpha}^* = \Sigma (X_i - \mu)/\nu$ で与えられ、その分散は $\alpha^2/\nu n$ である。 μ が未知のとき、 α の推定量の漸近分散はこれより小さくはなつ得ないから $\nu < 2$ のときその漸近分散は $\alpha^2/2n$ より大きくなるからいなければならない。実際 $\nu < 2$ のときは $\hat{\mu}^* = \min X_i$ とするとその分散の order は n^{-1} より小さくなるから、 $\hat{\alpha}^* = \Sigma (X_i - \hat{\mu}^*)/\nu$ の漸近分散は、 μ が既知の場合のそれに一致する。従つてモーメント法の漸近効率 $\nu < 2$ のとき $2/(\nu+3)$ に等しくなる。 $\nu = 2$ の場合には状況はより微妙になるが、 α, μ の最尤推定量 $\hat{\alpha}^*, \hat{\mu}^*$ とすると $\sqrt{n \log n}(\hat{\mu}^* - \mu), \sqrt{n}(\hat{\alpha}^* - \alpha)$ が漸近的に正規分布に従い、かつ $\hat{\alpha}^*$ の漸近分散は μ 既知の場合のそれに一致する。ことが示される。

これは2つの母数のうち、1つに關して正則條件が成立しているとき、そのことがもう一つの母数の推定量の漸近効率にどのように影響するかを示す例であるといふことができる。

Fisher は更に進んで3つの母数がすべて未知である場合を考へている。この場合モーメント法による ν の推定量は

$$4/\nu = M_3/M_2^{3/2}$$

で与えられる, μ の漸近分散は

$$6\nu(\nu+1)(\nu+5)/n$$

となる. 他方情報行列は $\nu > 2$ ならば

$$\frac{n}{\alpha^2} \begin{pmatrix} \frac{1}{\nu-2} & 1 & \frac{\alpha}{\nu-1} \\ 1 & \nu & \alpha \\ \frac{\alpha}{\nu-1} & \alpha & \alpha^2 \frac{d^2}{d\nu^2} \log \Gamma(\nu) \end{pmatrix}$$

とあるので ν の最尤推定量の漸近分散は

$$\frac{1}{n} \frac{1}{\frac{d^2}{d\nu^2} \log \Gamma(\nu) - \frac{1}{\nu-1} + \frac{1}{2(\nu-1)^2}}$$

となり, ν が大ならば, この値はほぼ

$$\frac{6}{n} (\nu-1) (\nu^2 - 2\nu + \frac{6}{5})$$

となる, 従って $\nu \rightarrow \infty$ ならばモーメント法の漸近効率¹は 1 に近づく, しかし ν が小さいときには, 与えられた小さいとして, Fisher のいろいろな ν に対応する相対効率²の値を計算している. しかしここでだけだけ $\nu < 2$ の場合まで同じ式で計算してしまっているが, これは勿論正しくない, この場合 μ の推定量の分散の order が n^{-1} よりも小さくなることに注意しているが, (しかしこのことは ν の推定に及ぼす影

響については、正しく認識して、なおったまじに思われる。

正しくは、この場合次のまじにある。 μ を既知とする。 α μ の同時推定において、情報行列は

$$n \begin{pmatrix} \nu/\alpha^2 & 1/\alpha \\ 1/\alpha & \frac{d^2}{d\nu^2} \log \Gamma(\nu) \end{pmatrix}$$

となるから、 ν の最尤推定量の漸近分散は

$$\frac{1}{n} \frac{1}{\frac{d^2}{d\nu^2} \log \Gamma(\nu) - \frac{1}{\nu}}$$

となる。 μ が未知の場合にも $\nu \leq 2$ ならば、 ν の最尤推定量の漸近分散はこの式から計算される値に一致することゝ証明できる。

更に Fisher はこの論文の中で Pearson 系の分布型で、かつ位置・尺度が未知の分布についての母数推定を論じているが、その中には分布の両端が切れている非正則の場合もよくまねてゐる。よくは一様分布について、その分布の中心の推定量として

$$\hat{\mu}^* = (\min X_i + \max X_i)/2$$

を取る。 $n(\hat{\mu}^* - \mu)$ が漸近的に両側指数分布に従うこと、すなわち分散の order は n^{-2} になること、他方 $\hat{\mu}^0 = \bar{X}$ の分散は n^{-1} に下がること、従つてその漸近効率はい 0 になること

とを言っているが、この際 $\hat{\theta}_n$ の「良さ」については特に何も言っていない。

Fisher は 1925 年の論文で、最尤推定量が漸近的に有効であることを示したのみならず、更にこの 2 次の漸近効率についても論じている。いま X_1, X_2, \dots, X_n が互いに独立に同じ分布に従うとき、この情報量を $I(\theta)$ 、一対 $\hat{\theta}_n = \hat{\theta}_n(X_1, \dots, X_n)$ を推定量、 $\hat{\theta}_n$ の分布の持つ情報量を $I_{\hat{\theta}_n}$ とするとき、つねに

$$I_{\hat{\theta}_n} \leq n I(\theta)$$

であるが $n \rightarrow \infty$ のとき

$$I_{\hat{\theta}_n} / n I \rightarrow 1$$

となるとき $\hat{\theta}_n$ は (一次の) 漸近有効性を持つという。これがふつうの意味の漸近有効性に一致することは証明できる。

ところで Fisher は更に 2 次の漸近効率を考え、

$$\lim_{n \rightarrow \infty} n I - I_{\hat{\theta}_n}$$

を $\hat{\theta}_n$ の漸近情報量損失と呼んだ。そうして正則の場合には漸近情報量損失が

$$I(\theta) \nabla \left\{ \frac{\partial^2}{\partial \theta^2} \log f(X, \theta) \mid \frac{\partial}{\partial \theta} \log f(X, \theta) \right\}$$

で与えられることを示し、かつ、これが一次漸近有効推定量の中で最小の情報量損失を与えることを示している。この四り

の Fisher の議論は極りと不明確であるが、この大筋の結論は誤っていない。その論理は後に Rao に与えられて明確にされたものである。

ところでこのように最尤推定量の二次しきものは Pfanzagl, Takeuchi-Akaiwa の議論によれば 3 次の有効性があり立つためには、密度関数が θ に関して 2 回微分可能であることが必要である。しかし 2 次微分可能性がたり立たないとき、最尤推定量の二次の漸近効率に関する上記の結論がたり立たないことを、Fisher は次の例で示している。

いま X_1, \dots, X_n が互いに独立に両側指数分布:

$$f(x, \theta) = \frac{1}{2} \exp -|x - \theta|$$

に従うとすると、 θ の最尤推定量は $\hat{\theta} = \text{med } X_i$ となる。

しかしこの場合

$$\frac{\partial}{\partial \theta} \log f(x - \theta) = -\text{sgn}(x - \theta)$$

であるから、情報量は $I = 1$ であり、 $\sqrt{n}(\hat{\theta} - \theta)$ は漸近的に平均の分散 1 の正規分布に従うから、 $\hat{\theta}$ は一次の漸近有効性を持っている。ところでこの場合 $\hat{\theta}$ の情報量損失は

$$l_n = V\left(\frac{\partial}{\partial \theta} L(\hat{\theta})\right)$$

で与えられるが、ここで $\partial L / \partial \theta = -\sum \text{sgn } |X_i - \theta|$ であるこ

と仮定する。

$$l_n = 2 \sqrt{\# X_i < \theta \mid \text{med } X_i}$$

これは具体的に次の式で与えられることが知られる。

$$n - \frac{n(n+1)}{n-3} \left\{ 1 - \frac{(n-1)!}{\left(\frac{n-1}{2}!\right)^2 2^{n-2}} \right\}$$

より、Fisher はこの量が漸近的に

$$\sqrt{n/\pi} - 4$$

に近づくこと、すなわちその大きさが正則の場合と異なり、 $n^{1/2}$ の order に近づくことを示した。Fisher はこの例と他の二つの論文でもくわしく取り上げ引用している。

この場合、 $\hat{\theta}^*$ は漸近情報量損失に関して二次の有効性を持つ。すなわち、

$$\lim_{n \rightarrow \infty} \frac{1}{\sqrt{n}} (\bar{n} I - I \hat{\theta}_n)$$

を最小にするものでないことを示すことができる。しかしこの場合 $\hat{\theta}_n^*$ の漸近分散が $1/(n - l_n)$ に近づくことを示す(他方

$$\hat{\theta}_n^{**} = \frac{1}{2} \{ X[\frac{n}{2} - t\sqrt{n}] + X[\frac{n}{2} + t\sqrt{n}] \}$$

として、 t を適当に選べば、この漸近分散が $1/(n - c_n)$ という形で与えられる。かつ $\lim (l_n - c_n)/\sqrt{n} > 0$ となるように選ぶことができる。より正確に言えば $\hat{\theta}_n^{**}$ の

情報量については $I_{\hat{\theta}_n} \geq 1/V(\hat{\theta}_n) = n - c_n$ となる

$$\lim_{n \rightarrow \infty} \frac{1}{n} (I_{\hat{\theta}_n} - I_{\theta_0}) > 0$$

となることを示さねばならない。

このことは Takeuchi-Akahira [3] に示されている。この例については最大推定量 $\hat{\theta}_n$ の意味の 2 次漸近有効性を持つこと、つまり事実に対応するものであることを示す。

Fisher は非正則な場合の議論においても、鋭い直観と深い洞察力を持って、このことを知らしめたのである。

References

- [1] Akahira, M. and Takeuchi, K. (1981). Asymptotic Efficiency of Statistical Estimators: Concepts and Higher Order Asymptotic Efficiency. Lecture Notes in Statistics 7, Springer, New York.
- [2] Fisher, R.A. (1950). Contributions to Mathematical Statistics. Wiley, New York.
- [3] Takeuchi, K. and Akahira, M. (1976). On the second order asymptotic efficiencies of estimators. Proceedings of the Third Japan-USSR Symposium on Probability Theory, Lecture Notes in Mathematics 550, 608-638, Springer, Berlin.
- [4] Woodroffe, M. (1972). Maximum likelihood estimation of a translation parameter of a truncated distribution. Ann. Math. Statist., 43, 113-122.